

本周周报（5.25-5.31）

刘昊南

本周工作

1. 本周对 MongoDB 的存储做了优化，并对 PostgreSQL 做了测试
2. MongoDB 因为没有固定的 schema，所以每个 document 的每个字段都必须记录字段名，从而带来了额外的存储开销，用出租车数据对此做了测试，插入了 100 万条记录
 - a) 使用原先的 document 的字段名（plateNumber, location, time, isPassengerIn, speed, direction）进行存储，数据消耗的空间为 230M，索引消耗的空间为 120M，总计 370M
 - b) 将 document 的字段名改为单个字母（n、loc、t、p、v、d）后，再进行插入，数据消耗的空间为 110M，索引消耗空间为 120M
 - c) 改变 document 的字段名后，数据消耗的空间减少了一半，而索引消耗的空间保持不变，这说明 MongoDB 在存储数值型数据时，字段占用的空间可能超过了数值占用的空间，从而带来了很大的开销，而在存储文本数据时，字段名占用的空间相对与文本则很小
3. 学习了 PostgreSQL+PostGis 的使用，在自己的主机上搭建起了 PostgreSQL 数据库并安装了 PostGis 插件，使用 PostgreSQL 的 JDBC Driver 和 PostGis 的 jar 包编写了数据插入的程序，对 PostgreSQL 做了测试，table 的 schema 如下

栏位	资料表 "public.taxis"	型别	修饰词
plate_number	:	character(8)	:
location	:	geography(Point,4326)	:
time	:	timestamp without time zone	:
is_passenger_in	:	boolean	:
speed	:	smallint	:
direction	:	smallint	:

索引:

```
"taxi_location_idx" gist (location)
"taxi_plate_number_idx" btree (plate_number)
"taxi_time_idx" btree ("time")
```

4. 对 PostgreSQL 使用出租车数据测试的结果
 - a) 数据量：100 万条
 - b) 插入数据消耗时间：58 秒
 - c) 建立索引消耗时间：27 秒
 - d) table 消耗的空间：110M
 - e) 索引消耗的空间：135M

f) 消耗总空间：250M

下周计划

1. 根据 MongoDB 和 PostgreSQL 的测试结果，在对 MongoDB 做过优化后，二者消耗的空间十分接近，MongoDB 的索引消耗空间还要略少，但是以 100 万条数据消耗 250M 的存储效率推算，3 亿条数据仍要消耗 75G 的磁盘空间
2. 之所以会比二进制文件（总共 13G）消耗的空间大，目前总结起来有几个原因：
 - a) 二进制文件里每条记录都没有车牌号，因为车牌号是文件名，文件里的所有记录都是属于这个车牌号的，省去了许多空间，而在数据库中我们不可能为每个车牌号建一个表，所以每条记录都必须包含车牌号的字符串
 - b) 二进制文件里时间戳采用 8 个字节记录，但是数据库里 timestamp 要消耗 30 个字节
 - c) 二进制文件里经纬坐标使用两个 float 记录，而 PostGis 里包含单个 Point 的 geometry 要消耗 32 个字节
 - d) 数据库里需要建立索引，从而带来了额外的开销
3. 准备用更大的数据集（5000 万条）对 MongoDB 和 PostgreSQL 做比较测试，测试插入速度、空间性能和查询性能